Task 7:

Collect and organize your data

Deliverables for this Task

1) A single Excel or SPSS spreadsheet with all of the data that you have collected, and a key that explains the meanings of your abbreviations or codes;

2) Your Methods section, revised to change everything to *past tense*, and including any last-minute changes or observations about the data collection and data coding methods, which you will update after collecting your data.

Step by step

Overview

One of the biggest advantages of experimental research is that the data collection is designed specifically to optimize the relevance and quality of the data for a specific set of hypotheses. You don't have to rely on other researchers' data and explain why it *might* be relevant for your hypotheses. Instead, you go out and collect your own data.

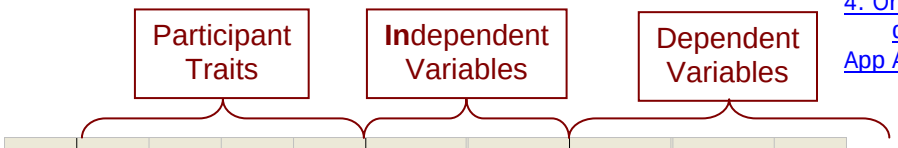
This Task focuses on your data collection activities. Data collection is actually quite complex, and you've already completed most of the process. It makes no sense to collect data without a research problem, in particular a research problem that you can show is important and interesting. It's self-defeating to collect data without planning what you will collect and how you will do it – without proper planning, the data collected is most often useless. And you're simply not allowed to collect data without permission if you work in some organization. You've completed all of these steps already by putting together your Research Proposal.

There are basically four steps left to complete your data collection:

1. prepare all materials and equipment;
2. recruit participants;
3. "run" the participants in your experimental conditions;
4. organize and code the data to prepare it for analysis;

Sample Deliverables

This is what your data spreadsheet should look like:



Participant number	Gender	Major	Age (years)	First Language	Music Exposure (Between)	Reading Order (Within)	Narrative Text # Correct (in 10)	Technical Text # Correct (in 10)	
1	1	2	18	1	1	1	8	10	
2	1	2	22	1	1	2	8	10	
3	2	2	20	1	1	1	10	10	
6	2	2	19	1	1	2	9	7	
7	2	2	28	1	1	1	9	9	
8	1	1	19	2	1	1	10	10	
9	2	2	24	1	1	2	8	10	
10	1	2	21	2	1	1	9	10	
11	1	1	20	2	1	1	9	8	
12	1	1	20	2	1	2	9	9	
19	2	2	19	1	1	2	8	9	
22	1	2	19	1	1	2	9	9	
4	2	2	18	2	2	1	6	9	
5	2	1	20	2	2	2	6	8	
13	1	1	22	1	2	1	9	7	
14	1	1	20	1	2	1	6	10	
15	1	2	19	1	2	1	6	9	
16	2	2	30	1	2	2	10	10	
17	1	2	19	1	2	1	10	10	
18	2	1	24	1	2	1	8	10	
20	1	1	22	1	2	2	9	10	
21	2	1	22	2	2	1	8	10	
23	2	1	20	2	2	2	10	9	
24	1	1	20	2	2	2	8	9	

… (more dependent variables)…

Of course, you will also need a ***key*** on a separate sheet of paper, so that we all know what your numbers mean:

Gender (1 = Feminine; 2 = Masculine)

Major (1 = Psychology; 2 = Other)

First Language (1 = English; 2 = Other)

Music Exposure (1 = no music; 2 = with music)

Reading Order (1 = narrative, technical; 2 = technical, narrative)

Note:
All of the data about a given participant goes on **the same** line.

Collecting your data, step by step

1. Prepare all materials and equipment

Start your countdown to data collection by completing and checking off these items one by one.

- Prepare one master copy of everything that you will distribute to the participants;

First, however, review all instructions and written materials that the participants will use. Salkind's (2006) third Commandment of data collection is: "Make sure that the data collection form you are using is clear and easy to use." Make sure to check that all instructions are very easy *for the participants* to understand:

a) the instructions should be 9th grade reading level at most, if you're using university students (check this with the grammar checker in MS Word);

b) rephrase any pronouns except for "you";

c) make sure that no sentences are more than 15 words long;

d) make sure that you are not using any technical terms at all;

e) make sure that there are no passive sentences (ex: "The instructions will ***be read*** to you." should be: "You will hear the instructions." – "be read" is an example of a passive verb);

These are all ways to simplify the instructions and to ensure that the participants will understand them correctly.

- Make sure that you have all the equipment that you need;

Be absolutely sure that it all works! Test it at least three times. Be sure to have extra batteries, if you're using them. Check on the availability of backup equipment.

- Rehearse your procedure with your partner to make sure that everyone knows that to do when;

In past Research Methods classes, the students who rehearsed their procedures collected data without stress and without problems. The students who didn't rehearse their procedures had many more problems and found data collection much more stressful.

- Consider seriously the possibility of doing Pilot Testing;

A "dry run" or "dress rehearsal" of your experiment with only two or three people in each condition takes extra time and effort but is an excellent way to detect and prevent problems in your problems and materials (See Mitchell & Jolley, 2007, p. 530).

Observe very closely how your pilot participants react to instructions and materials. Do they hesitate? Do they have questions? Do they end up doing the wrong things? These are signs that they didn't understand what to do. Do they do things faster or slower than you expected? Maybe the task is easier or harder than you imagined.

These observations will allow you to revise the instructions and materials to make them more appropriate for your participants.

> Remember:
> Do NOT include your pilot data in your final data set!

- Don't count on being able to download materials from the internet when you run your experiment;

Leaving things to do for the last minute is an invitation for stress and mistakes to take over.

- Have your supervisor and/or an experienced experimenter double check your materials and procedure;

Word has it that frustrated student experimenters invented Murphy's Law ("Everything that can go wrong, will go wrong – at the worst possible moment."). Planning and preparation are the best antidotes.

- Make more copies than you actually need of your materials, just in case.

You might have more participants available than you expected. If you make extra copies of your materials, you may be able to collect extra data instead of missing the opportunity.

2. Recruit participants

You have three options for recruiting participants. Get informed about them and get organized for them as soon as possible. This is Salkind's (2006) seventh Commandment of data collection.

a) "Hunting and gathering" on your own. If you need participants from a special population, then you MUST have gotten official permission from the Institutional Review Board or Ethics Committee first!

Even with more usual participants, if you don't participate in Open Research Day or sign up to take advantage of the Participant Pool, then you're on your own. You have to hunt down and persuade (often one by one!) the participants that you need. This is very time consuming, so plan ahead and leave a lot of time for it.

Regardless of how you recruit your participants, no-shows are a fact of life. Don't count on everyone showing up when you want them to.

b) Open Research Day. Many Psychology Departments organize a day when the hundreds of Introduction to Psychology students can fulfill their requirement to participate in a psychology experiment. You're running an experiment, so you can sign up to recruit participants on this day.

You may have noticed, ahem!, that with undergraduates attendance is difficult to predict. So, it's hard to tell how many students will show up. Even if for some reason you can't get enough participants on Open Research Day it's a very convenient option.

Sign up to assure your participation! Arrive well ahead of time to set up and test your equipment. Think through how your procedure will work with larger groups of participants.

c) The Participant Pool. Similarly, most Psychology Departments have someone help the experimenters find participants. In this case, you sign up as an experimenter, post a time and place when you will need participants, and pray for them to come. Then you repeat the process until you have recruited enough participants.

Sometimes, this means that you will have to run the participants one by one. Be aware that this can take a lot of time, so plan ahead! Also, it's often the case that you will have to go out and persuade people to sign up.

Please note: Currently, access to Open Research Day and the Participant Pool is restricted to project that already have IRB approval. Therefore, these options are not open to us.

3. "Run" the participants [deliverables]

You've already done all the planning; you've reserved and tested the equipment; you've copied all the materials; you've got people signed up to come. It's show time!

> But, please plan ahead before you actually collect your data. At the very least, read to the end of this Task.

Salkind's (2006, p. 150) fifth Commandment of data collection contains a very relevant warning here: "Do not rely on other people to collect or transfer your data unless you personally have trained them and are confident that they understand the data collection process as well as you do." That's why you have done so much planning and have a script to follow: these can be used as training materials for your assistants. There's much too much at stake for you to let just anyone collect data for you: don't ever simply turn over your materials for someone else to "just" run your experiment.

Trials and sessions. Each time a participant responds to a stimulus or group of stimuli it's called a *trial*. If a participant reads one text and answers questions about it, that's one trial. If a participant sees 20 words and for each one has to give a response, that's 20 trials.

You will probably run more than one group of participants – each time you run a group of participants, it's called a *session*. Sometimes you can run one session for two of your experimental conditions; sometimes you run one session for each condition. Make sure that you know how many sessions you will need to run. That will help you figure out how long your data collection will take.

You want to focus on three things when you're running your participants: Ethics, Consistency, and Accuracy.

Ethics. It is your responsibility to treat all participants extremely ethically. This means, among other things, that it is your job as an experimenter to ensure that:

a) they participate with *informed consent* – they have to know what they're getting into and agree to it – and they can leave the experiment at any time;

b) they are protected as much as possible from physical, social, and psychological risks;

c) their participation and responses remain anonymous;

d) it's also a good idea to be nice to your participants – they're doing you a favor;

For anonymity, assign each participant a unique code like "12" for the 12th participant in the experiment. Get participants' names only when you really need them. Make sure that the participants' names are not visible during data coding – start using the participant codes as soon as possible, so that what you know about the participants doesn't affect how you code their responses.

Read the APA's Ethical Principles of Psychologists and Code of Conduct at http://www.apa.org/ethics/code2002.html for details. Michell & Jolley's (2007) Appendix C has a clear and helpful discussion of ethical considerations.

Consistency. Your main goal when running the participants is to ensure that ALL of them participate under the same conditions, except for the differences that are planned parts of the

experiment. But things happen -- there may be a flash of lightning, a car crash, a telephone rings, someone starts singing an opera, etc. – that may affect the outcome.

If this happens, don't panic. Continue the experiment, if possible. What most experimenters do is just take notes during each run of the experiment. They observe anything that was non-routine: papers fell on the floor so it took longer to collect the stimuli; a door slammed during the instructions; there was a short brown-out; etc. These notes may be helpful in explaining the results later on.

Random assignment. Part of consistency is random assignment. Don't forget to assign participants to the different experimental conditions *randomly*. Use a list of random numbers: the first 15 random numbers are the participants in the first experimental condition; the second 15 random numbers are the participants in the second experimental condition, etc. Then sort the numbers numerically: the first number tells you which condition is for the first participant who shows up, the second number the condition for the second participant, etc.

In practice, researchers emphasize *random assignment* (of participants to experimental conditions) instead of random selection because random assignment is more practical to carry out. Perhaps the most common technique for random assignment is to use a table of random numbers like the one in Appendix A.

Say that you need to assign 40 participants to 4 experimental conditions. You can follow these steps:

- Choose a two-digit number arbitrarily (from the serial number of a dollar bill, a license plate, the page of a book opened at random, etc.).
- Use the first digit and the column number and the second digit as the row number to find where to start in your table of random numbers.
- Go down the column one number at a time. If it is less than 41, then write that number on the list for experimental condition #1 until you have enough participants for condition 1. If the number is more than 40 (your total) or was already seen, then skip it. If you get to the bottom of a column, go to the next column.
- Repeat until you fill all of the experimental conditions.

Why is this random? For one thing, the assignment does not depend on any characteristics of the participants or the experimenter. If you did the sampling yourself, you might unconsciously assign the smart (or male, or white, or …) participants to one condition or other. For another, the random number table is designed to provide a sequence that's very close to pure randomness.

Accuracy. Accuracy in this case refers to being extra, super sure that participant 7 (for example) gave *these* responses to the background questionnaire, participated in *this* experimental condition with *those* materials, gave *these* other responses to the main task, etc. If any of this information gets mixed up, then the data goes to the trash! You have to be totally, absolutely sure. This is why researchers are usually required to keep their data for at least five years after the results have been published – so that their accuracy can be verified later on. This is Salkind's (2006) ninth Commandment of data collection: don't throw out the original data.

Collect your data. Go for it! Have the participants sit down and then follow your script as precisely as possible. Repeat until you have enough participants.

4. Organize and code the data

Congratulations! Now you have a big stack of papers or measurements – real data!

Make backup copies. If you collected measurements with a computer or other instruments, then the very, very first thing to do is back everything up:

1. Make an extra, backup copy of all of the data files,
2. Double-check that everything was copied correctly,
3. Make sure that the backup is write protected (can't be erased), and
4. Hide the backup copies someplace safe.

This is Salkind's (2006) fourth Commandment of data collection.

Get organized. Now make sure that:

a) all of the data for each participant is together in the same place,

b) you're absolutely sure that each piece is from *that* participant, and

c) each piece has *that* participant's code on it.

Reality check: if someone (crazed supervisor, mischievous child, enraged significant other, …) threw all of your data out the window, could you put it all back together again? If not, then your data isn't organized yet.

BTW, yes, you *can* write on your participants' response sheets. It's *your* data, as long as you don't change their responses. You may want to write group numbers, analysis codes, or preliminary counts on each sheet.

Coding your data

Your next step is to code the data: the information from the background questionnaire, the factors or independent variables, and the responses – your dependent measures – to make it easier to manipulate in the statistical analysis programs.

Data coding for dependent measures. The next step is to *score* or *code* the data. What score will *this* participant get for each of the things that you want to measure?

> Be sure to *score* or *code* the data in the exact same way for each and every participant in all of your conditions.
>
> *Don't* try to make up a coding procedure as you go.

Consistency is why you really need written instructions to follow and not just your intuition: if you follow explicit instructions then you will be more consistent. With written instructions, you can get help to do the coding and it's likely (not guaranteed!) that the different coders will almost do the same thing as you would.

In the easiest case, you will have a list of acceptable or "correct" responses and you will just count up the acceptable responses for each participant. Say you have a test questionnaire with 20 multiple-choice reading comprehension questions. For each question, you can simply check the participant's response again your list of "correct" responses. This is an easy method, but it doesn't provide a lot of information.

Advanced Example: coding recall of propositions. In more interesting cases of coding, there are some similarity judgments involved. For example, you may opt to identify "idea units" or *propositions* – units of information in a target text that participants were supposed to read and recall. It's absolutely normal for each participant to recall the same information in a different form – using different words and sentence structures. Actually, that's what makes it interesting to study.

Take an example where the sentence in the original text was this:

The young man went ashore to his house and made a fire.

Participant 6 wrote: *The warrior went home and sat by the fire.*

How many propositions did Participant 6 recall?

Figure 1. The analysis of a participant's response against the original propositions.

Original propositions	**Participant's response: *The warrior went home and sat by the fire.***
1 Man, young	Omitted
2 Man, go, ashore	Similar: *go* is the same; *man* is like *warrior*
3 Man, go, house	More similar: *go* is the same; *man* is like *warrior; house* is like *home*
4 House, his	Similar: *home* is like *his house*
5 (man), make, fire	Only *fire* is similar
6 And, go, make	*Go* and *and* are the same

Now what? Only bits and pieces are the same or similar. Which propositions can we say were recalled? This is exactly why you have to think the analysis through *before* you start analyzing your data.

Proposition 4. If we allow for synonyms, then *his house* and *home* will count as synonyms. So, proposition 4 can reasonably be counted as recalled – the same ideas are involved.

Proposition 3. For proposition 3, *go* is the same and we just decided that *his house* and *home* are the same. That's two out of three pieces of the proposition. *Warrior* is related to the original *man* but not a synonym: *warrior* is a kind of *man*. The substitution suggests that Participant 6 probably knows this fact and used it during comprehension. The change suggests that it wasn't a case of simple memorization – the reader is *thinking* about what he read, associating the text with what he knows.

For coding, one option is to say that *warrior* and *man* are close enough and conclude that proposition 3 was recalled. We're counting synonyms, subordinate terms (elements), and probably superordinates (categories) as equivalent for purposes of coding, under this option.

We can also say that this is a different *kind* of recall and count it separately. Call it *recall with inference*, for example, to show that the reader's not just remembering, but doing some additional mental acrobatics.

So far we have proposition 3 was *recalled* and proposition 4 was *recalled with inference.*

Proposition 6. Proposition 6 is the closest match left: two out of three parts of the proposition are the same. But the third item (*make* in the original, *sit* in the response) doesn't match at all; they're not even close. Again, we have different coding options. We can ignore it because all three parts of the proposition didn't show up in the response. Or we can say that it's recalled because two out of three is close enough. **There's no correct answer here: you have to decide on a clear criterion and use it systematically.** Another option is to count it as a third kind of recall, one in which the reader tried to link something from the text with additional knowledge (like *recall with inference*), but didn't link to anything directly related to the text. Call it *inference*, for example, to show that the reader's thinking, but that it's not so directly related to the text.

If you go for the last option, then propositions 5 and 6 get counted as inferences.

Proposition 2. Proposition 1 simply was omitted by the participant. That leaves proposition 2. Two out of three parts are the same as the original, and the third part is unrelated. This looks like another case of *inference* and it would make sense to code it that way. Another option is to ignore it on the grounds that proposition 3 was already coded and contains all of the information that showed up (*man* and *go*), so why give the reader "double credit" for it?

There's no correct answer here: you have to decide on clear criteria and use them systematically.

Don't forget to describe the details of your coding procedure in your Methods section.

If you do your coding with all three response types, then you have three dependent variables for each participant: the total propositions recalled, the total recalled with inference, the total with inference. This means you'll have three columns for dependent variables in your spreadsheet. You don't need all three (or to invent more) – this is just an example. It is one way that the coding is done in the research literature on text comprehension.

Data coding for background questionnaires and factors. You have to do something similar with the information from the background questionnaires. At least, you should do it for the parts of the questionnaire that you might want to test during analysis.

First you will need to devise a ***key*** so that we all know what your numbers mean, as in the examples below. We're treating each question on the questionnaire as if it were a factor and the different answers can be seen as levels of that factor. The coding, at least, is done in exactly the same way.

Gender (1 = Feminine; 2 = Masculine)

Major (1 = Psychology; 2 = Other)

First Language (1 = English; 2 = Other)

In this way, you enter the codes into the spreadsheet.

7 2 1 1

This example means that this participant (#7) was a male psychology major who's first language is English.

Finally, you will do the same thing for your factors / independent variables: give each level a code and enter the codes for each participant in the spreadsheet.

Triple check everything that you put into your spreadsheet. Errors of data entry make the results very difficult or impossible to interpret. Another advantage of working with a partner is that you have someone to help your review data entry.

Further Resources on Collecting your Data

In fact, there's surprisingly little information on actually collecting data in most Research Methods textbooks. They're usually focused more on theoretical aspects of research.

Mitchell, M. & Jolley, J. 2005/2007. *Research Design Explained [Appendix C]*. Belmont, CA: Thompson Wadsworth.

Salkind, N. 2006. *Exploring Research [Ch. 7]*. Upper Saddle River, NJ: Prentice-Hall.

Appendix A

Table of Random Numbers

39634 62349 74088 65564 16379 19713 39153 69459 17986 24537
14595 35050 40469 27478 44526 67331 93365 54526 22356 93208
30734 71571 83722 79712 25775 65178 07763 82928 31131 30196
64628 89126 91254 24090 25752 03091 39411 73146 06089 15630
42831 95113 43511 42082 15140 34733 68076 18292 69486 80468

80583 70361 41047 26792 78466 03395 17635 09697 82447 31405
00209 90404 99457 72570 42194 49043 24330 14939 09865 45906
05409 20830 01911 60767 55248 79253 12317 84120 77772 50103
95836 22530 91785 80210 34361 52228 33869 94332 83868 61672
65358 70469 87149 89509 72176 18103 55169 79954 72002 20582

72249 04037 36192 40221 14918 53437 60571 40995 55006 10694
41692 40581 93050 48734 34652 41577 04631 49184 39295 81776
61885 50796 96822 82002 07973 52925 75467 86013 98072 91942
48917 48129 48624 48248 91465 54898 61220 18721 67387 66575
88378 84299 12193 03785 49314 39761 99132 28775 45276 91816

77800 25734 09801 92087 02955 12872 89848 48579 06028 13827
24028 03405 01178 06316 81916 40170 53665 87202 88638 47121
86558 84750 43994 01760 96205 27937 45416 71964 52261 30781
78545 49201 05329 14182 10971 90472 44682 39304 19819 55799
14969 64623 82780 35686 30941 14622 04126 25498 95452 63937

58697 31973 06303 94202 62287 56164 79157 98375 24558 99241
38449 46438 91579 01907 72146 05764 22400 94490 49833 09258
62134 87244 73348 80114 78490 64735 31010 66975 28652 36166
72749 13347 65030 26128 49067 27904 49953 74674 94617 13317
81638 36566 42709 33717 59943 12027 46547 61303 46699 76243

46574 79670 10342 89543 75030 23428 29541 32501 89422 87474
11873 57196 32209 67663 07990 12288 59245 83638 23642 61715
13862 72778 09949 23096 01791 19472 14634 31690 36602 62943
08312 27886 82321 28666 72998 22514 51054 22940 31842 54245
11071 44430 94664 91294 35163 05494 32882 23904 41340 61185

82509 11842 86963 50307 07510 32545 90717 46856 86079 13769
07426 67341 80314 58910 93948 85738 69444 09370 58194 28207
57696 25592 91221 95386 15857 84645 89659 80535 93233 82798
08074 89810 48521 90740 02687 83117 74920 25954 99629 78978
20128 53721 01518 40699 20849 04710 38989 91322 56057 58573

00190 27157 83208 79446 92987 61357 38752 55424 94518 45205
23798 55425 32454 34611 39605 39981 74691 40836 30812 38563
85306 57995 68222 39055 43890 36956 84861 63624 04961 55439
99719 36036 74274 53901 34643 06157 89500 57514 93977 42403
95970 81452 48873 00784 58347 40269 11880 43395 28249 38743

56651 91460 92462 98566 72062 18556 55052 47614 80044 60015
71499 80220 35750 67337 47556 55272 55249 79100 34014 17037
66660 78443 47545 70736 65419 77489 70831 73237 14970 23129
35483 84563 79956 88618 54619 24853 59783 47537 88822 47227
09262 25041 57862 19203 86103 02800 23198 70639 43757 52064